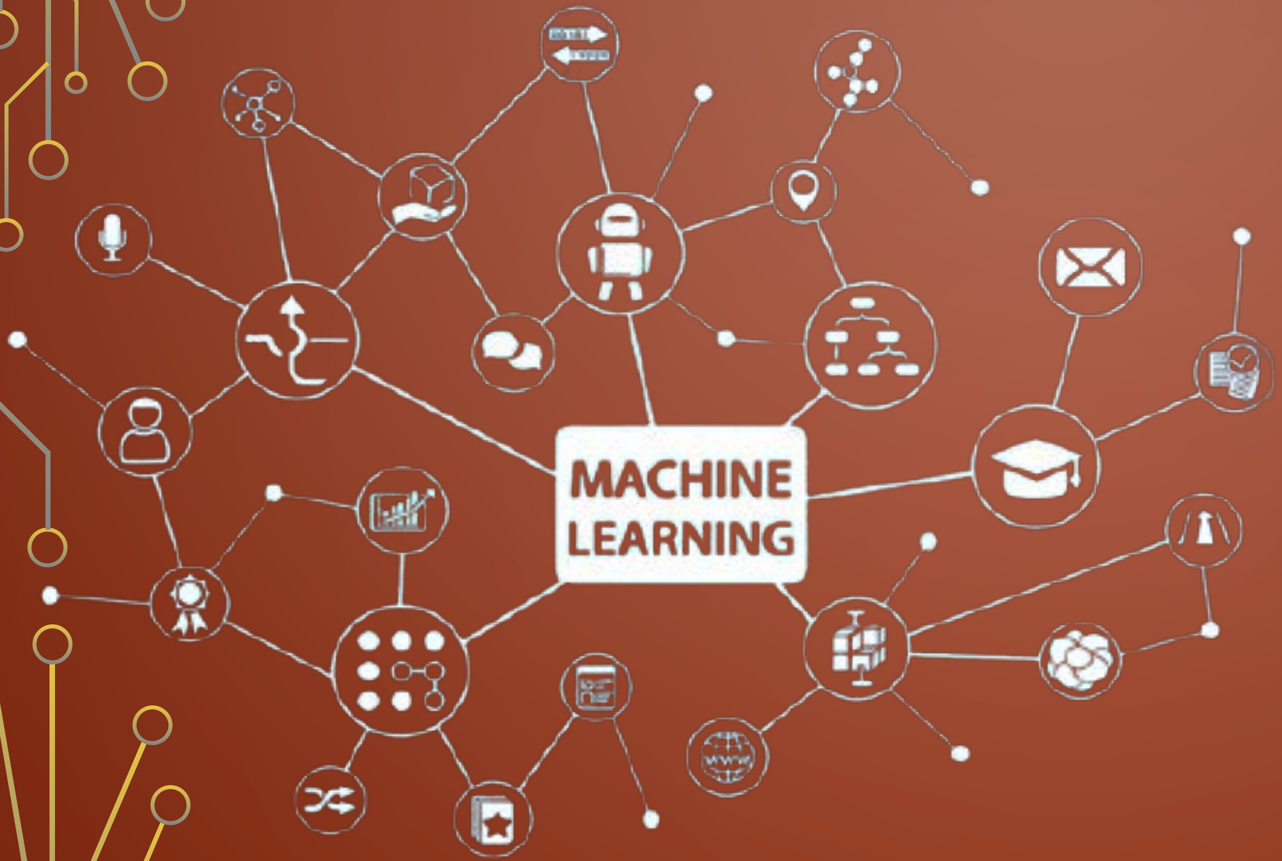
Talent · Technology · Solutions
Bellintegrator

Intelligent Agent
Neuton



# MACHINE LEARNING MADE SIMPLE

Suram Saraswati Anugna

Talent · Technology · Solutions

Bellintegrator

AI

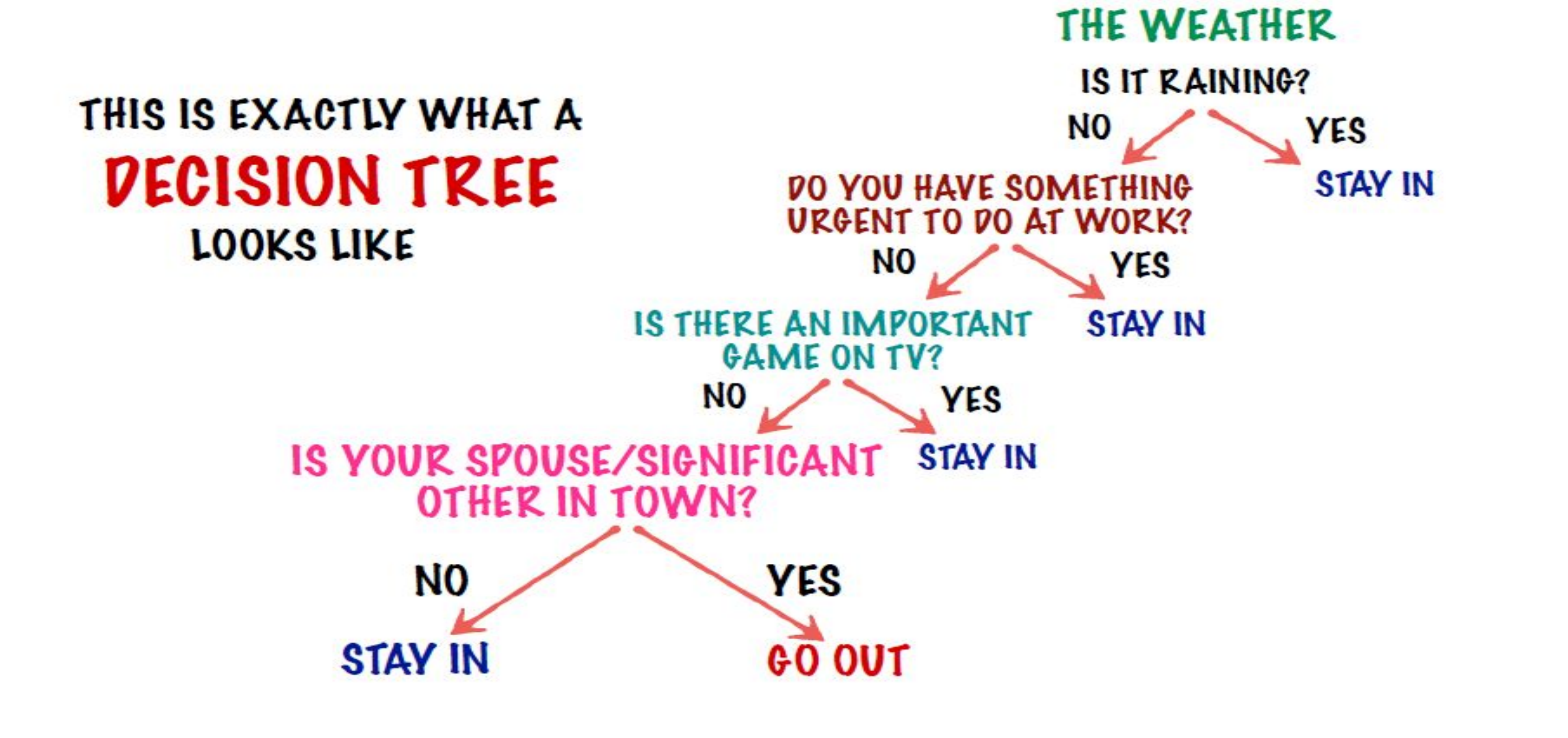
Intelligent Agent

Neuton

# Decision Tree Algorithm

# EXAMPLE

- What are some good predictors of whether you will stay in or go out for this weekend?
- Factors : Weather, your spouse in town, important game on TV, do you have something urgent to do at work.
- All these factors usually have some order of importance.



THIS IS EXACTLY WHAT A **DECISION TREE** LOOKS LIKE

THE WEATHER

IS IT RAINING?

NO

YES

STAY IN

DO YOU HAVE SOMETHING URGENT TO DO AT WORK?

NO

YES

STAY IN

IS THERE AN IMPORTANT GAME ON TV?

NO

YES

STAY IN

IS YOUR SPOUSE/SIGNIFICANT OTHER IN TOWN?

NO

YES

STAY IN

GO OUT

# WHAT ARE DECISION TREES?

- In Machine Learning decision tree predicts the outcome given the values of the input variables.
- It’s a Supervised learning algorithm.
- Decision Tree helps in making decision in structured manner and visually represents all the variables involved and the consequences in each case.

# Advantages and Disadvantages

1. Easy to interpret and perfect for visual representation.

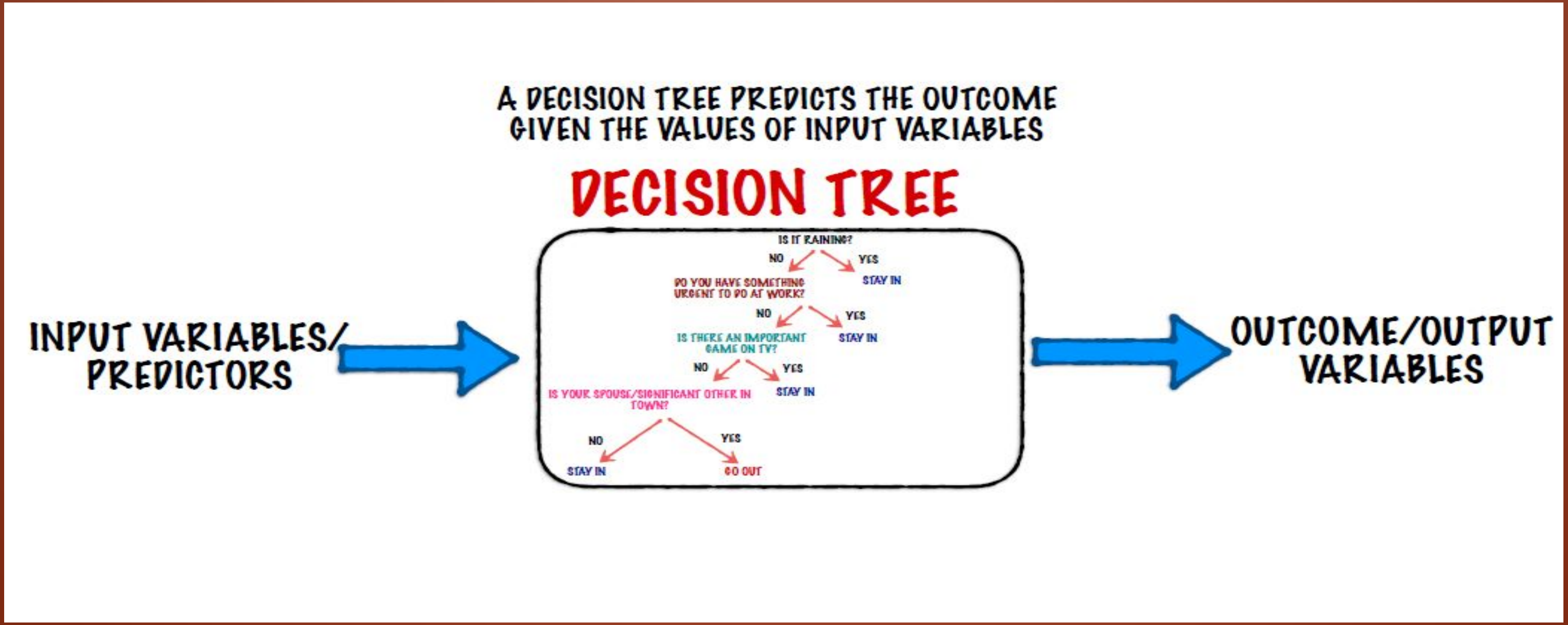
2. It works both with categorical and numerical features.

3. Nonlinear parameters don’t affect its performance.

4. Little effort is required in data preparations.

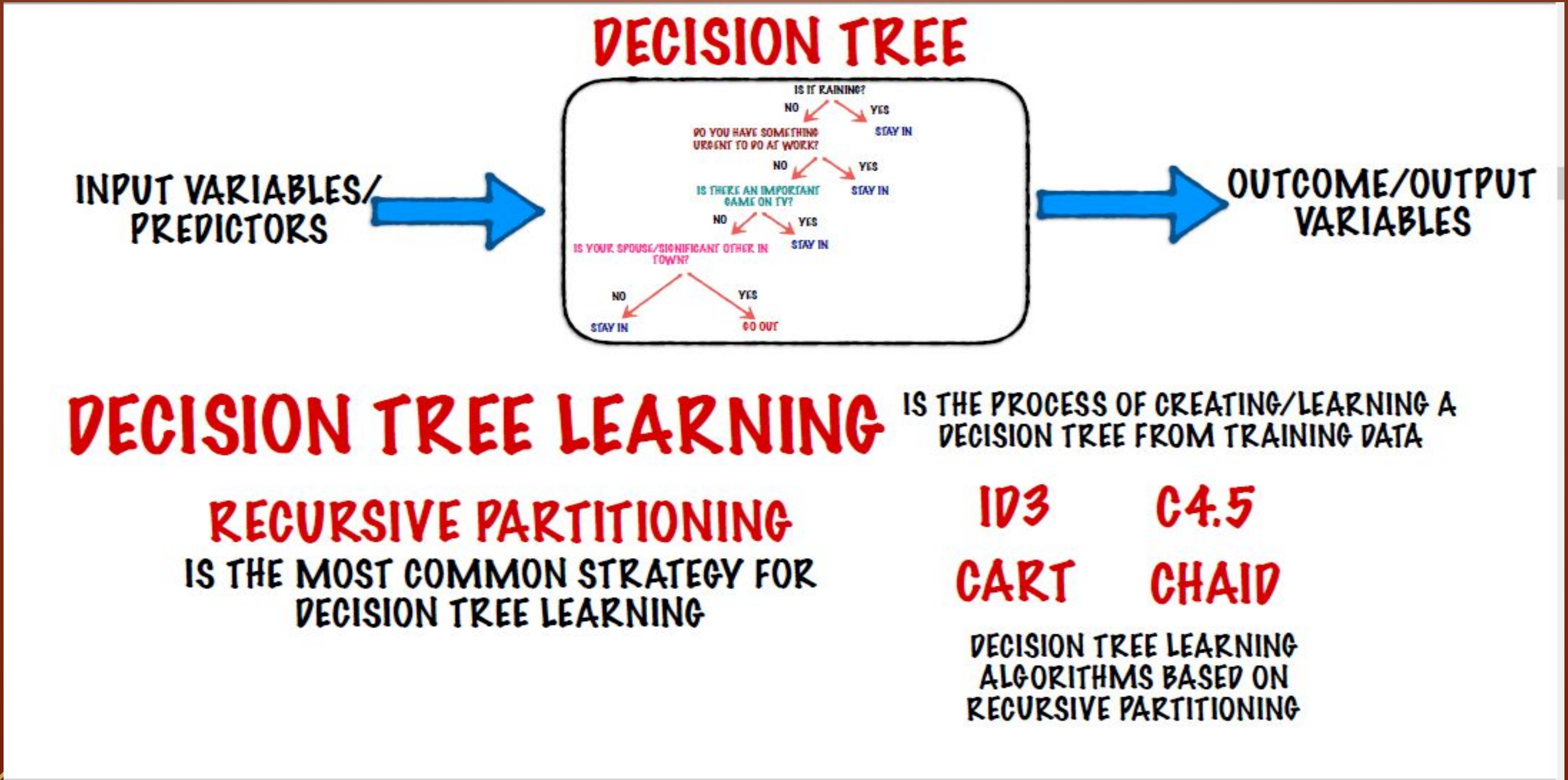
1. Overfitting as algorithm tends to pick up the noise in the data.
2. High Variance as model tends to get unstable with little variation in data.
3. Low bias which makes the model difficult to work with new data.

# WORKFLOW OF DECISION TREES



A DECISION TREE PREDICTS THE OUTCOME GIVEN THE VALUES OF INPUT VARIABLES

DECISION TREE
INPUT VARIABLES/ PREDICTORS
IS IT RAINING?
NO
YES
STAY IN
DO YOU HAVE SOMETHING URGENT TO DO AT WORK?
NO
YES
STAY IN
IS THERE AN IMPORTANT GAME ON TV?
NO
YES
STAY IN
IS YOUR SPOUSE/SIGNIFICANT OTHER IN TOWN?
NO
YES
STAY IN
GO OUT
OUTCOME/OUTPUT VARIABLES



# DECISION TREE

INPUT VARIABLES/ PREDICTORS → → OUTCOME/OUTPUT VARIABLES

IS IT RAINING?
NO
YES
STAY IN
DO YOU HAVE SOMETHING URGENT TO DO AT WORK?
NO
YES
STAY IN
IS THERE AN IMPORTANT GAME ON TV?
NO
YES
STAY IN
IS YOUR SPOUSE/SIGNIFICANT OTHER IN TOWN?
NO
YES
STAY IN
GO OUT

**DECISION TREE LEARNING** IS THE PROCESS OF CREATING/LEARNING A DECISION TREE FROM TRAINING DATA

**RECURSIVE PARTITIONING**

IS THE MOST COMMON STRATEGY FOR DECISION TREE LEARNING

**ID3** **C4.5**

**CART** **CHAID**

DECISION TREE LEARNING ALGORITHMS BASED ON RECURSIVE PARTITIONING

## Keywords of Decision Tree

- The topmost feature would be called as **root node**
- Then we have other features as **branches**
- The endpoints are called **leaf nodes.**

**Entropy** : The measure of randomness and unpredictability in the dataset.

Range(0 to 1)

$$E = -\sum_{i}^{C} p_i \log_2 p_i$$

**Information Gain** is measure of entropy after the dataset has been split.

Range(0 to 1)

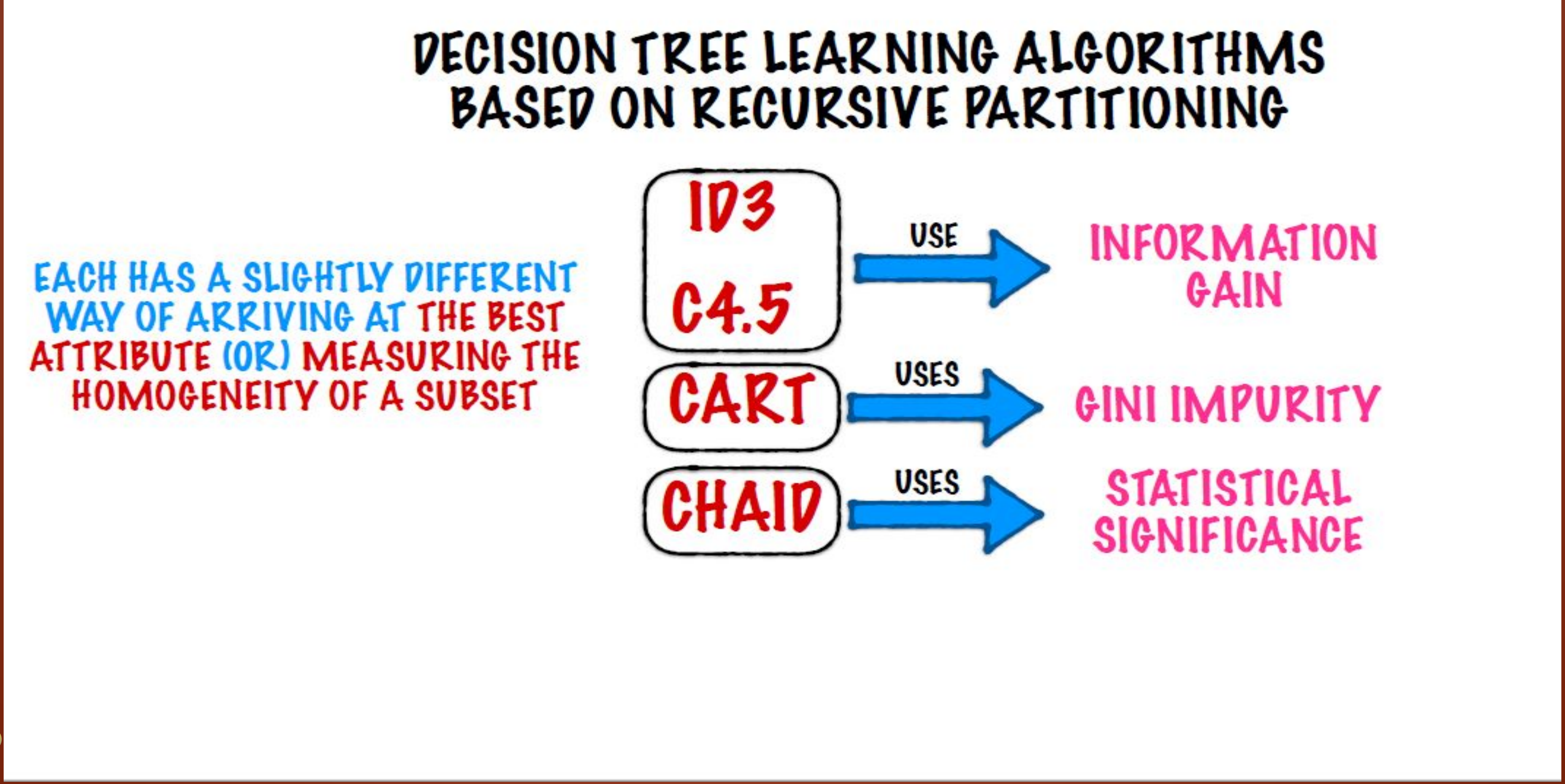
$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

**Gini Impurity** is used for calculation of purity of a split.

Range (0 to 1)

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

# How the Decision Tree works



# DECISION TREE LEARNING ALGORITHMS BASED ON RECURSIVE PARTITIONING

EACH HAS A SLIGHTLY DIFFERENT WAY OF ARRIVING AT THE BEST ATTRIBUTE (OR) MEASURING THE HOMOGENEITY OF A SUBSET

ID3

C4.5

USE

INFORMATION GAIN

CART

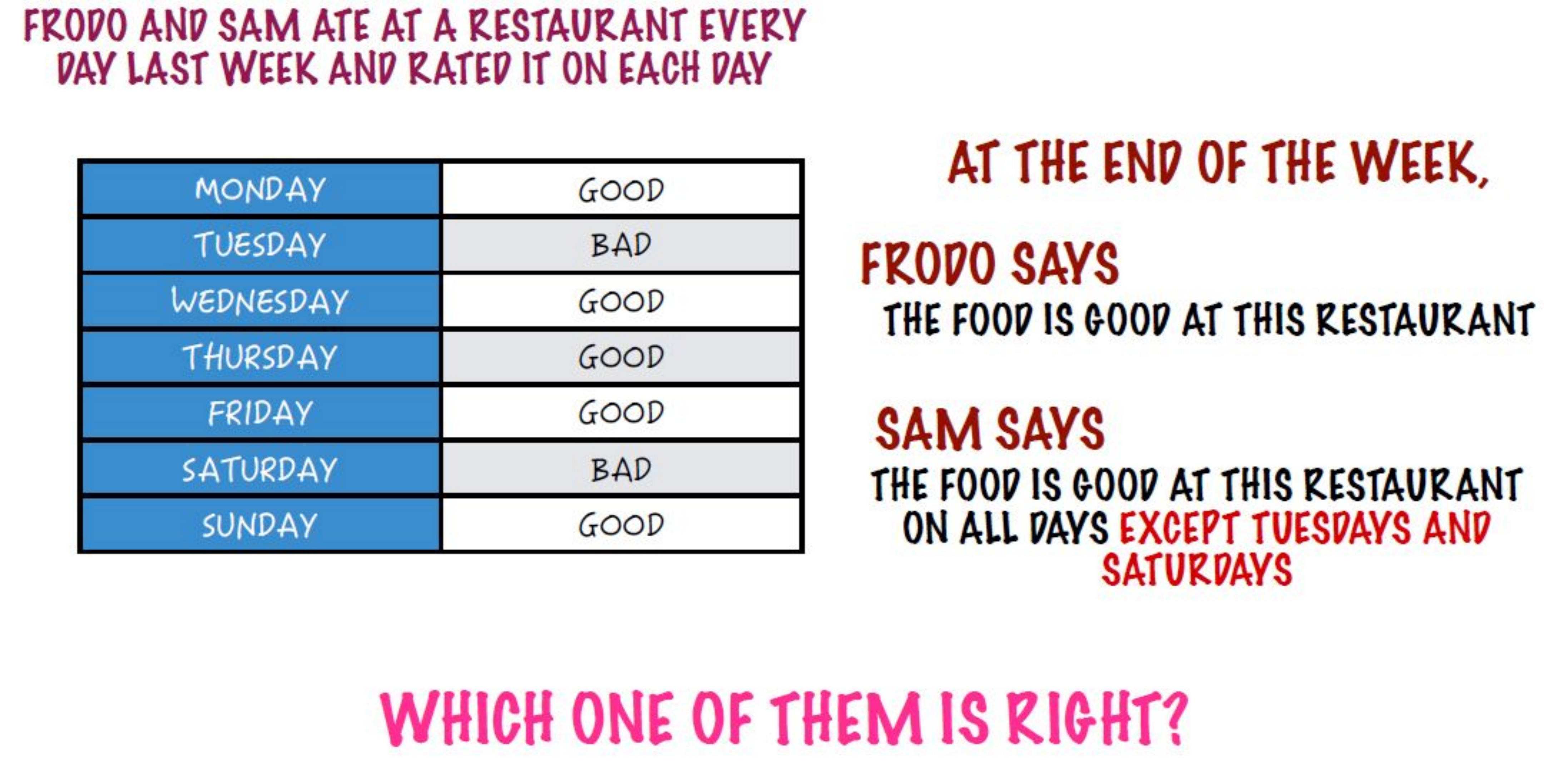
USES

GINI IMPURITY

CHAID

USES

STATISTICAL SIGNIFICANCE



FRODO AND SAM ATE AT A RESTAURANT EVERY DAY LAST WEEK AND RATED IT ON EACH DAY

| | |
|---|---|
| MONDAY | GOOD |
| TUESDAY | BAD |
| WEDNESDAY | GOOD |
| THURSDAY | GOOD |
| FRIDAY | GOOD |
| SATURDAY | BAD |
| SUNDAY | GOOD |

AT THE END OF THE WEEK,

FRODO SAYS

THE FOOD IS GOOD AT THIS RESTAURANT

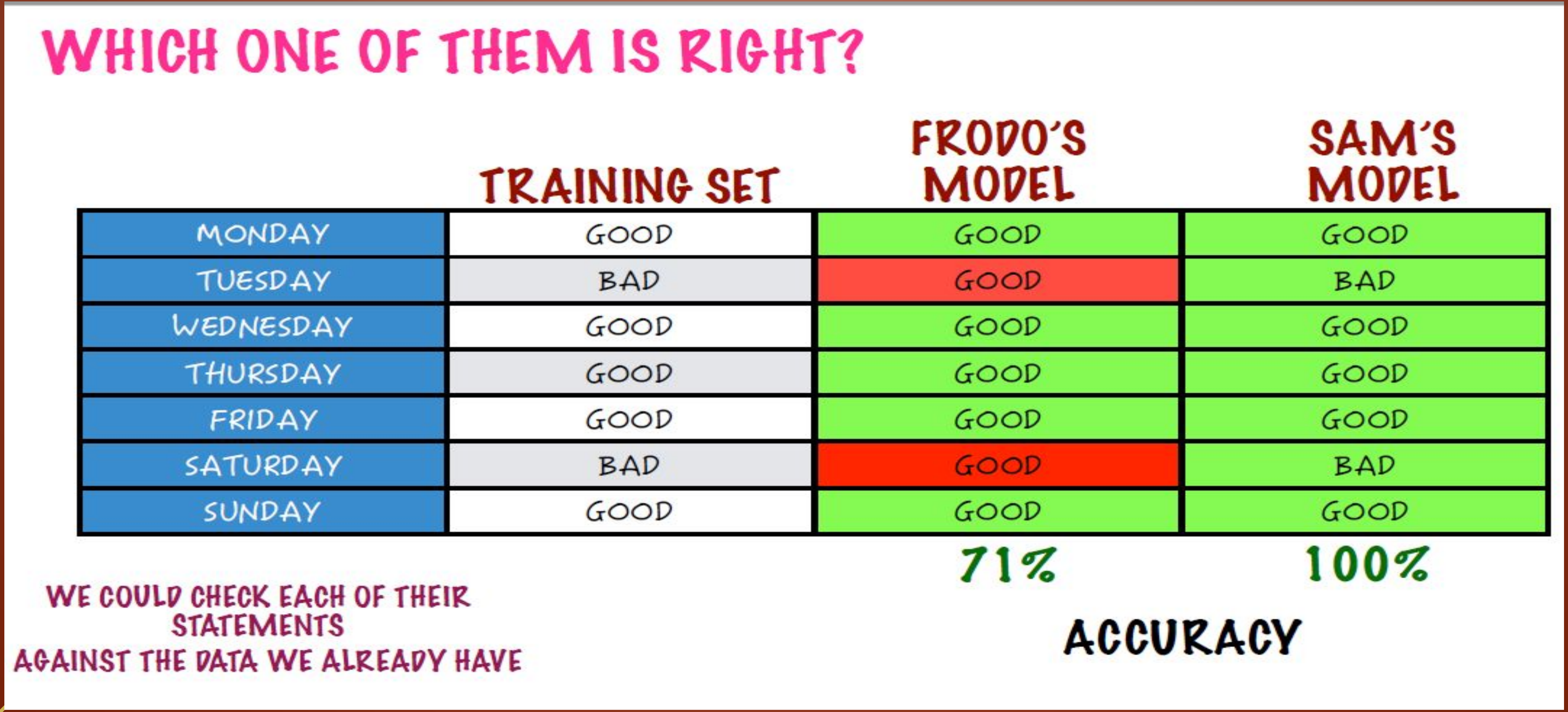
SAM SAYS

THE FOOD IS GOOD AT THIS RESTAURANT ON ALL DAYS EXCEPT TUESDAYS AND SATURDAYS

WHICH ONE OF THEM IS RIGHT?

Talent · Technology · Solutions
Bellintegrator

Intelligent Agent
Neuton

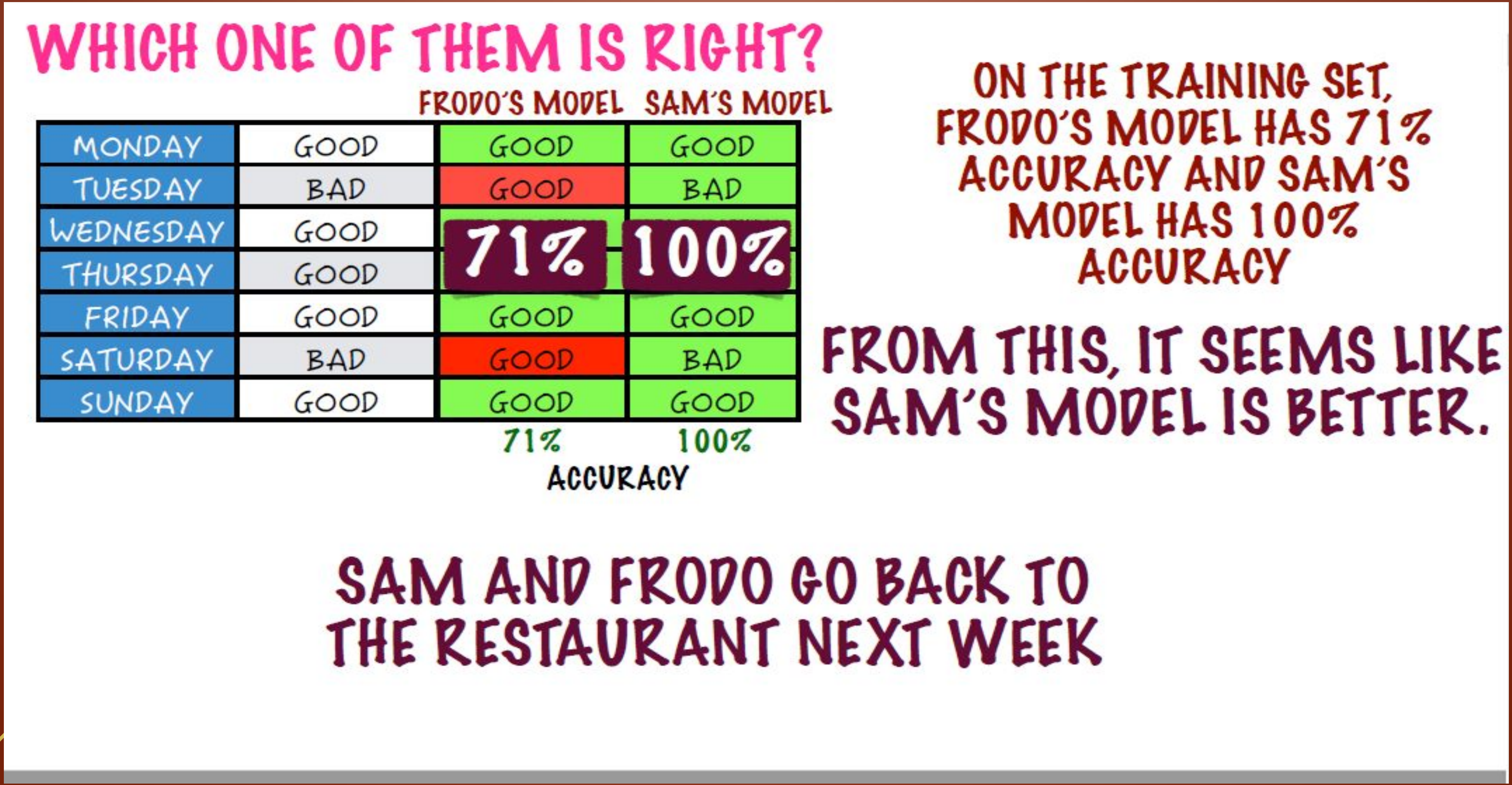


# WHICH ONE OF THEM IS RIGHT?

| | TRAINING SET | FRODO'S MODEL | SAM'S MODEL |
|---|---|---|---|
| MONDAY | GOOD | GOOD | GOOD |
| TUESDAY | BAD | GOOD | BAD |
| WEDNESDAY | GOOD | GOOD | GOOD |
| THURSDAY | GOOD | GOOD | GOOD |
| FRIDAY | GOOD | GOOD | GOOD |
| SATURDAY | BAD | GOOD | BAD |
| SUNDAY | GOOD | GOOD | GOOD |
| | | 71% | 100% |

ACCURACY

WE COULD CHECK EACH OF THEIR STATEMENTS
AGAINST THE DATA WE ALREADY HAVE



# WHICH ONE OF THEM IS RIGHT?

| | | FRODO'S MODEL | SAM'S MODEL |
|---|---|---|---|
| MONDAY | GOOD | GOOD | GOOD |
| TUESDAY | BAD | GOOD | BAD |
| WEDNESDAY | GOOD | 71% | 100% |
| THURSDAY | GOOD | | |
| FRIDAY | GOOD | GOOD | GOOD |
| SATURDAY | BAD | GOOD | BAD |
| SUNDAY | GOOD | GOOD | GOOD |
| | ACCURACY | 71% | 100% |

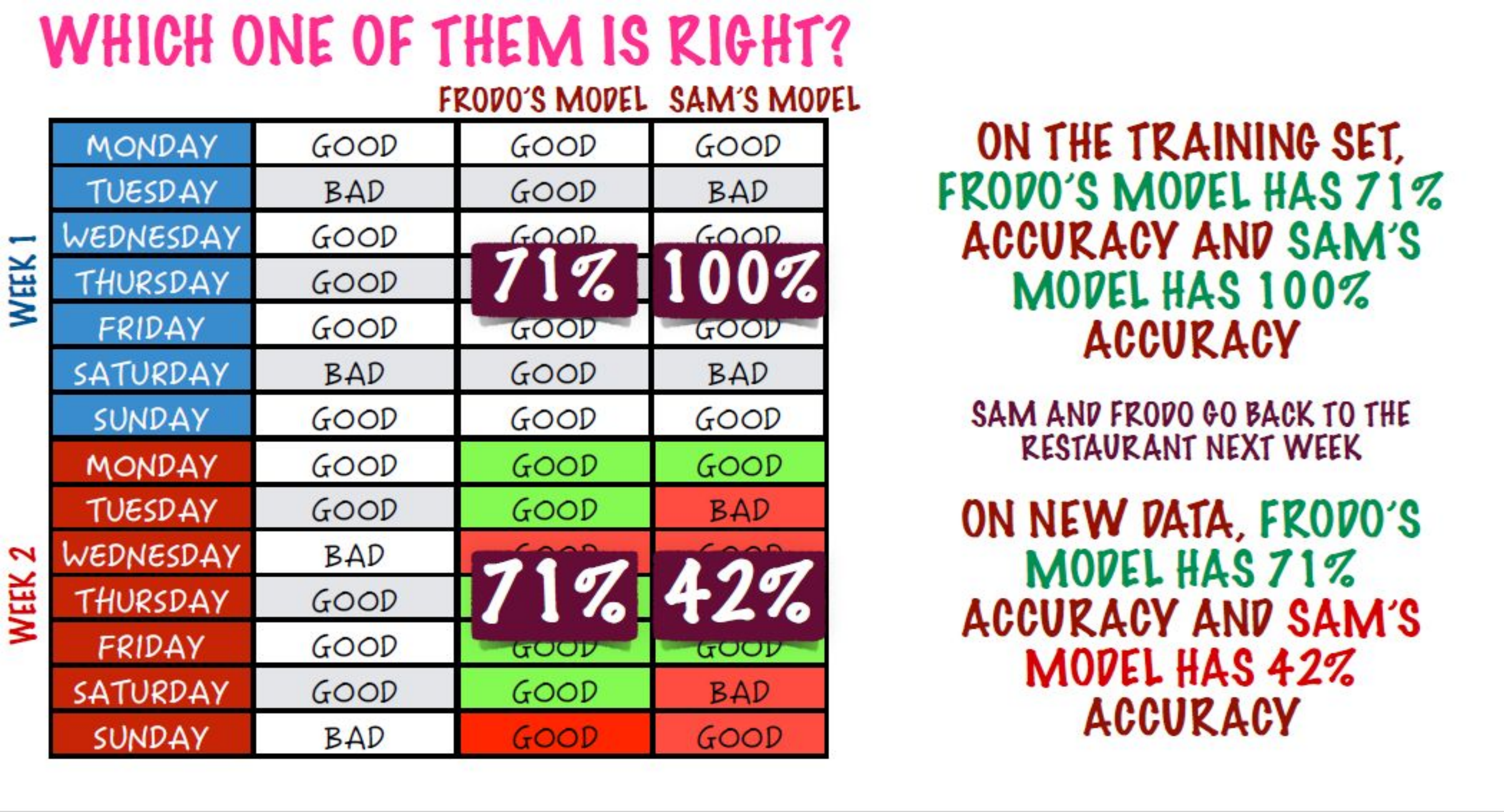
ON THE TRAINING SET, FRODO'S MODEL HAS 71% ACCURACY AND SAM'S MODEL HAS 100% ACCURACY

FROM THIS, IT SEEMS LIKE SAM'S MODEL IS BETTER.

SAM AND FRODO GO BACK TO THE RESTAURANT NEXT WEEK

Talent·Technology·Solutions Bellintegrator

Intelligent Agent Neuton



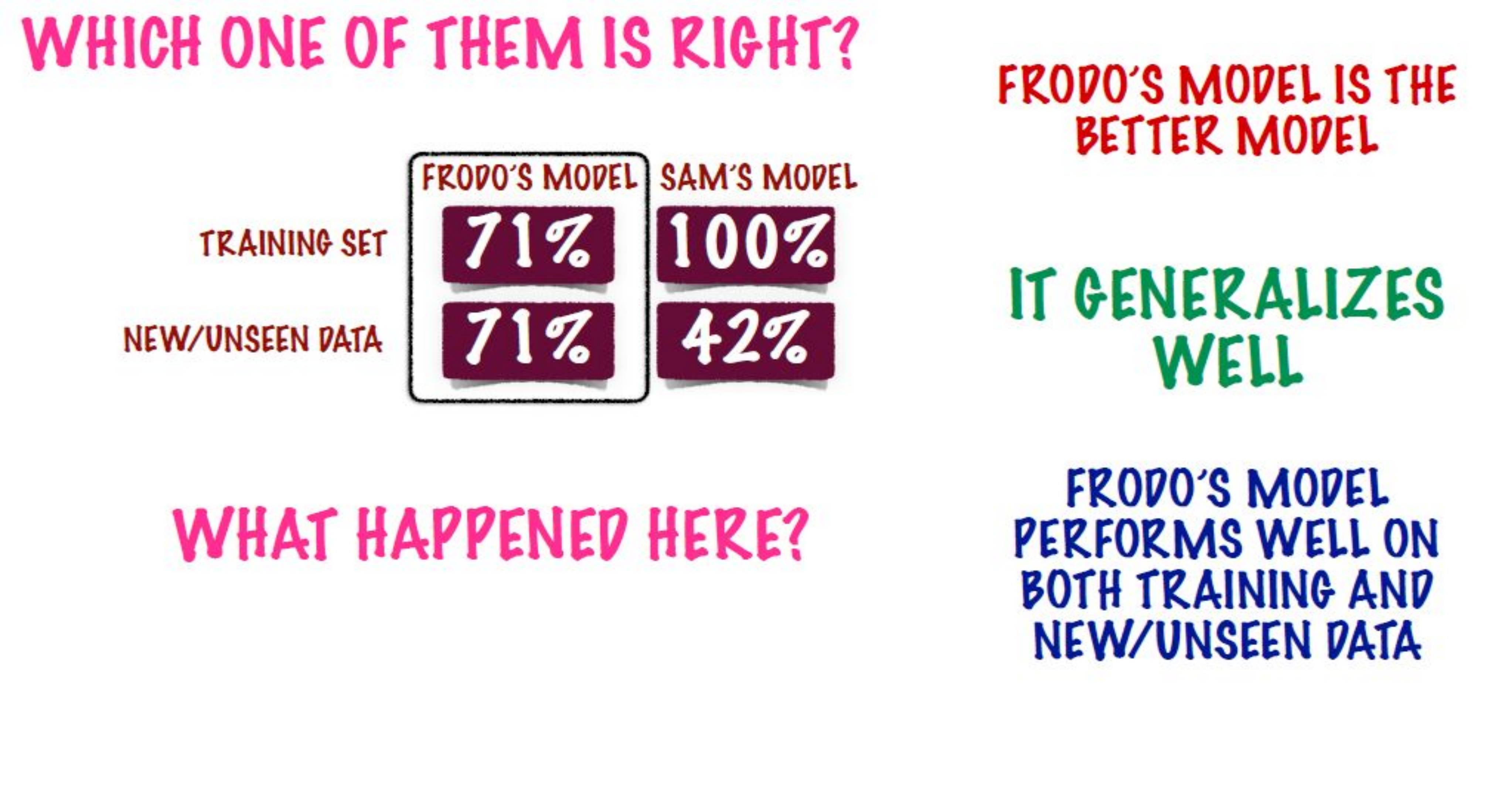
## WHICH ONE OF THEM IS RIGHT?

| | | | FRODO'S MODEL | SAM'S MODEL |
|---|---|---|---|---|
| WEEK 1 | MONDAY | GOOD | GOOD | GOOD |
| | TUESDAY | BAD | GOOD | BAD |
| | WEDNESDAY | GOOD | GOOD | GOOD |
| | THURSDAY | GOOD | 71% | 100% |
| | FRIDAY | GOOD | GOOD | GOOD |
| | SATURDAY | BAD | GOOD | BAD |
| | SUNDAY | GOOD | GOOD | GOOD |
| WEEK 2 | MONDAY | GOOD | GOOD | GOOD |
| | TUESDAY | GOOD | GOOD | BAD |
| | WEDNESDAY | BAD | GOOD | GOOD |
| | THURSDAY | GOOD | 71% | 42% |
| | FRIDAY | GOOD | GOOD | GOOD |
| | SATURDAY | GOOD | GOOD | BAD |
| | SUNDAY | BAD | GOOD | GOOD |

ON THE TRAINING SET, FRODO'S MODEL HAS 71% ACCURACY AND SAM'S MODEL HAS 100% ACCURACY

SAM AND FRODO GO BACK TO THE RESTAURANT NEXT WEEK

ON NEW DATA, FRODO'S MODEL HAS 71% ACCURACY AND SAM'S MODEL HAS 42% ACCURACY



## WHICH ONE OF THEM IS RIGHT?

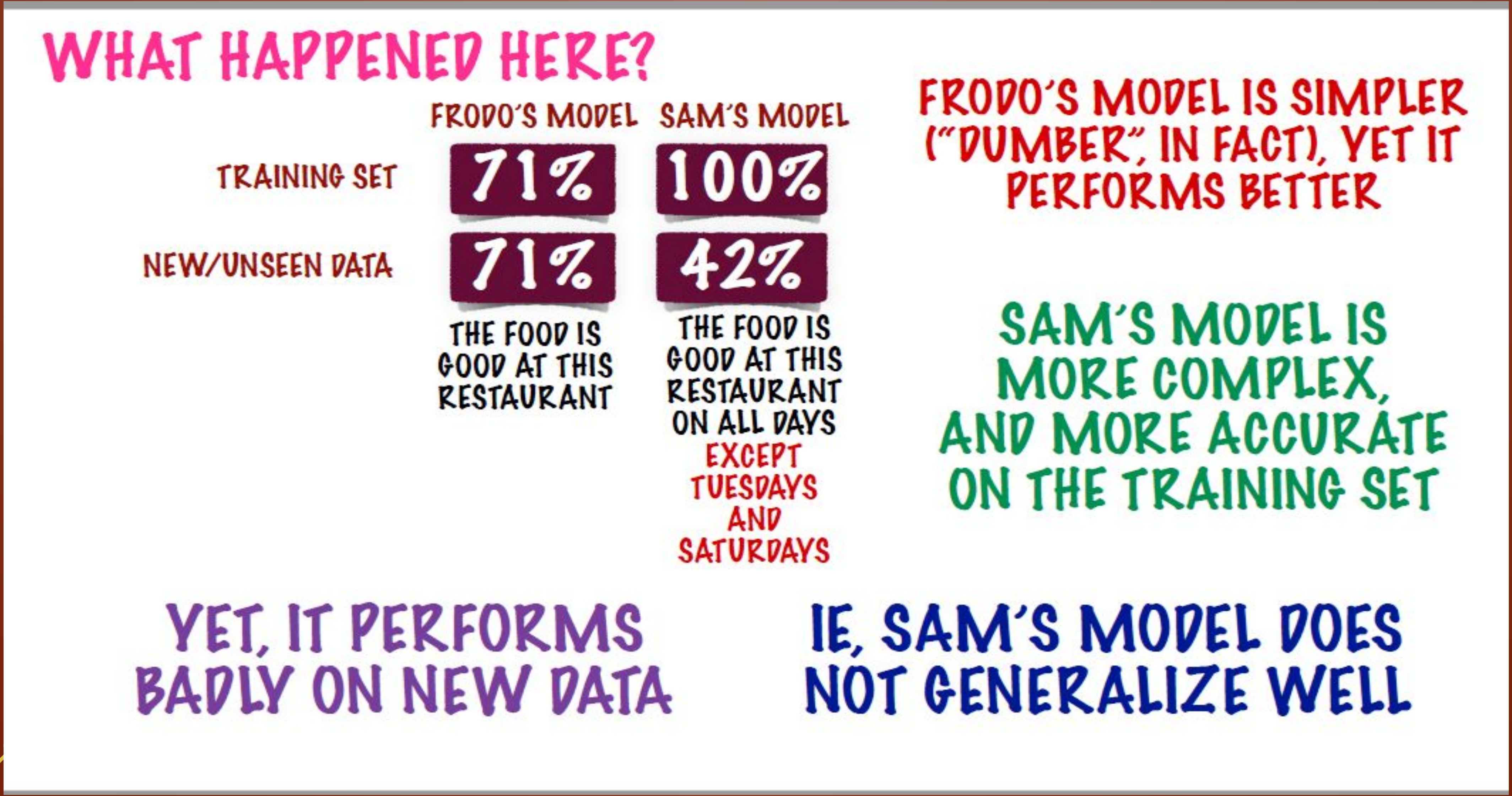
| | FRODO'S MODEL | SAM'S MODEL |
|---|---|---|
| TRAINING SET | 71% | 100% |
| NEW/UNSEEN DATA | 71% | 42% |

WHAT HAPPENED HERE?

FRODO'S MODEL IS THE BETTER MODEL

IT GENERALIZES WELL

FRODO'S MODEL PERFORMS WELL ON BOTH TRAINING AND NEW/UNSEEN DATA



# WHAT HAPPENED HERE?

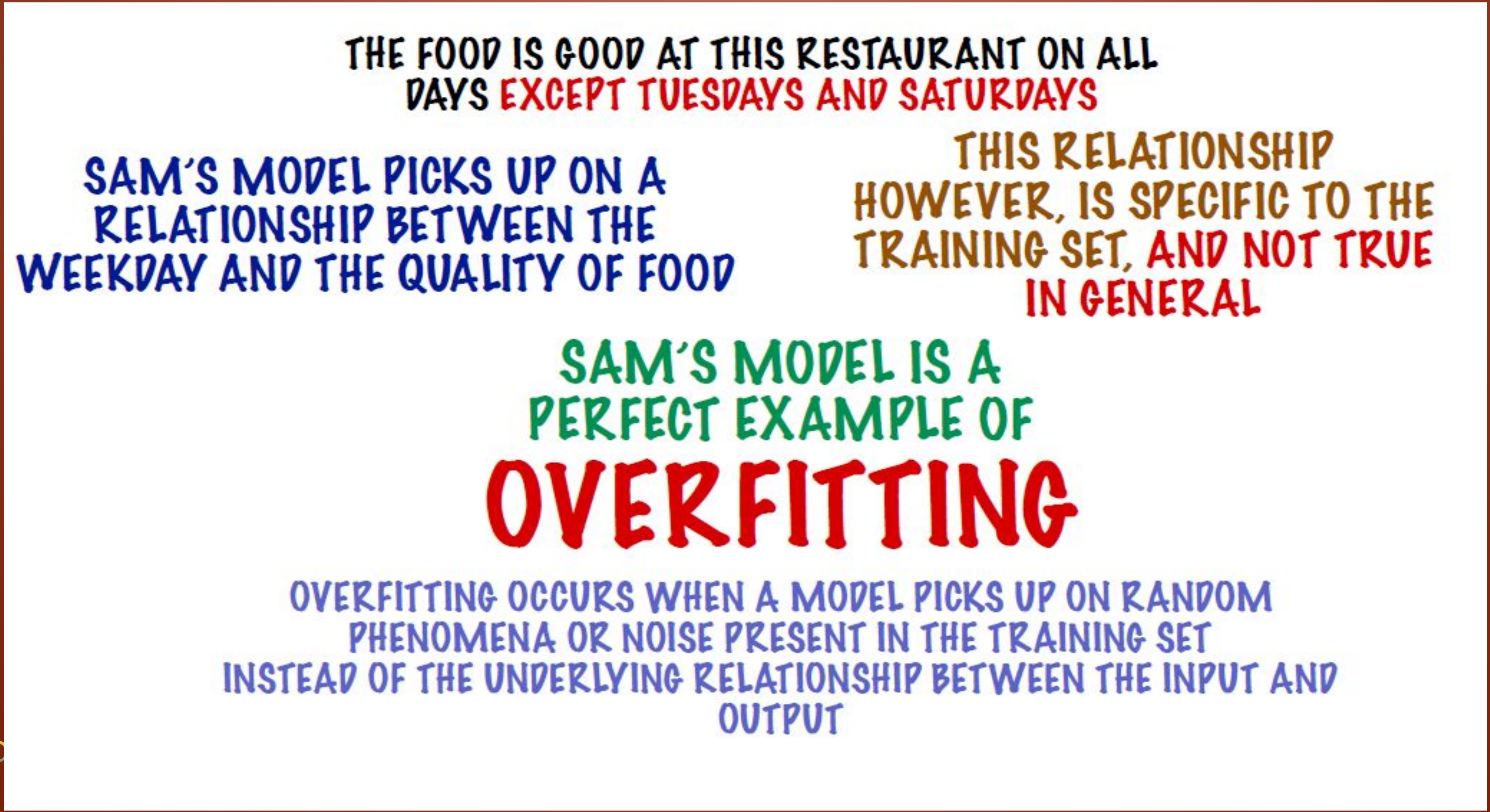
| | FRODO'S MODEL | SAM'S MODEL |
|---|---|---|
| TRAINING SET | 71% | 100% |
| NEW/UNSEEN DATA | 71% | 42% |
| | THE FOOD IS GOOD AT THIS RESTAURANT | THE FOOD IS GOOD AT THIS RESTAURANT ON ALL DAYS EXCEPT TUESDAYS AND SATURDAYS |

FRODO'S MODEL IS SIMPLER ("DUMBER", IN FACT), YET IT PERFORMS BETTER

SAM'S MODEL IS MORE COMPLEX, AND MORE ACCURATE ON THE TRAINING SET

YET, IT PERFORMS BADLY ON NEW DATA

IE, SAM'S MODEL DOES NOT GENERALIZE WELL



THE FOOD IS GOOD AT THIS RESTAURANT ON ALL DAYS EXCEPT TUESDAYS AND SATURDAYS

SAM'S MODEL PICKS UP ON A RELATIONSHIP BETWEEN THE WEEKDAY AND THE QUALITY OF FOOD

THIS RELATIONSHIP HOWEVER, IS SPECIFIC TO THE TRAINING SET, AND NOT TRUE IN GENERAL

SAM'S MODEL IS A PERFECT EXAMPLE OF

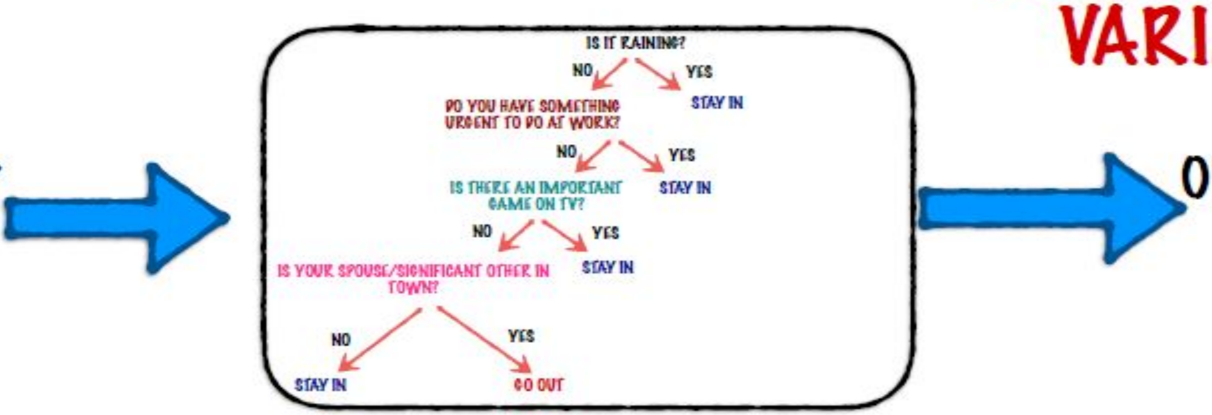
# OVERFITTING

OVERFITTING OCCURS WHEN A MODEL PICKS UP ON RANDOM PHENOMENA OR NOISE PRESENT IN THE TRAINING SET INSTEAD OF THE UNDERLYING RELATIONSHIP BETWEEN THE INPUT AND OUTPUT

A DECISION TREE CAN BE USED TO SOLVE MACHINE LEARNING PROBLEMS

A DECISION TREE PREDICTS THE OUTCOME GIVEN THE VALUES OF INPUT VARIABLES

INPUT VARIABLES/ PREDICTORS



IS IT RAINING?
NO
YES
STAY IN
DO YOU HAVE SOMETHING URGENT TO DO AT WORK?
NO
YES
STAY IN
IS THERE AN IMPORTANT GAME ON TV?
NO
YES
STAY IN
IS YOUR SPOUSE/SIGNIFICANT OTHER IN TOWN?
NO
YES
STAY IN
GO OUT

OUTCOME/OUTPUT VARIABLES

DECISION TREES ARE VERY PRONE TO THE RISK OF OVERFITTING

ENSEMBLE LEARNING CAN MITIGATE THE RISK OF OVERFITTING

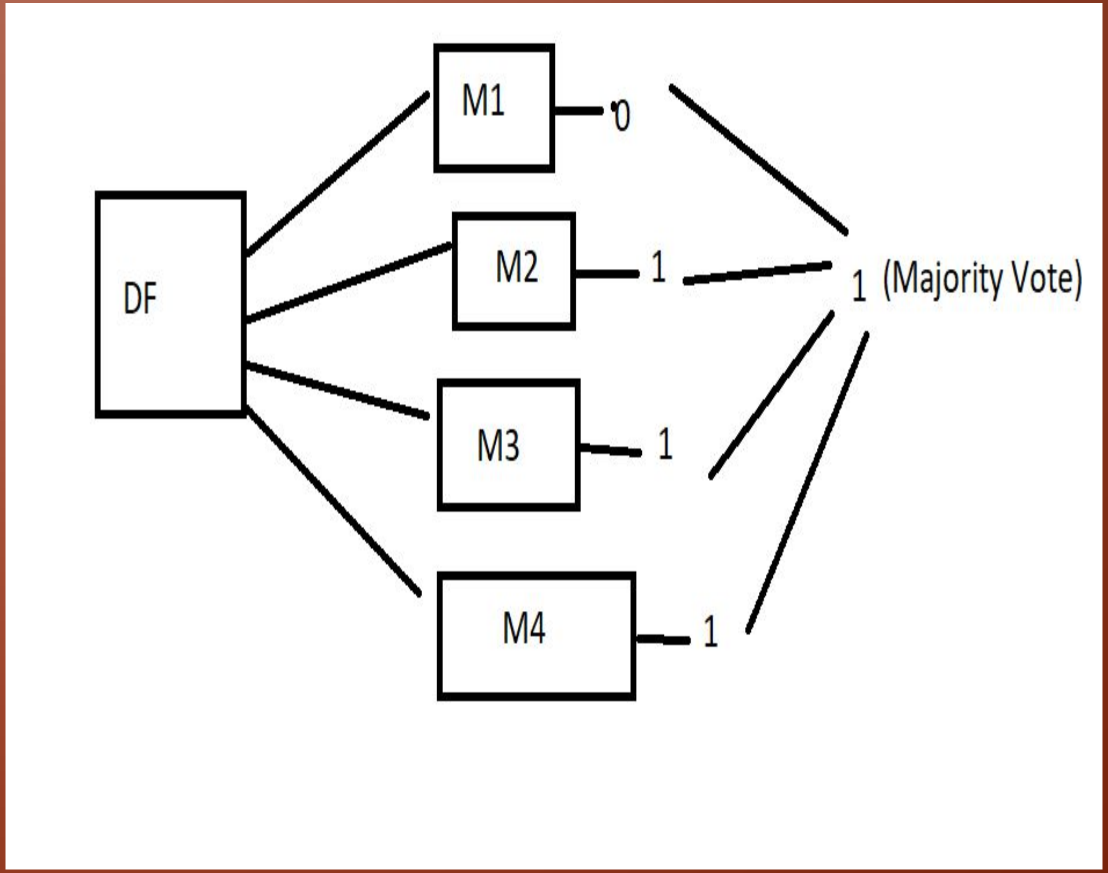
Talent · Technology · Solutions
Bellintegrator

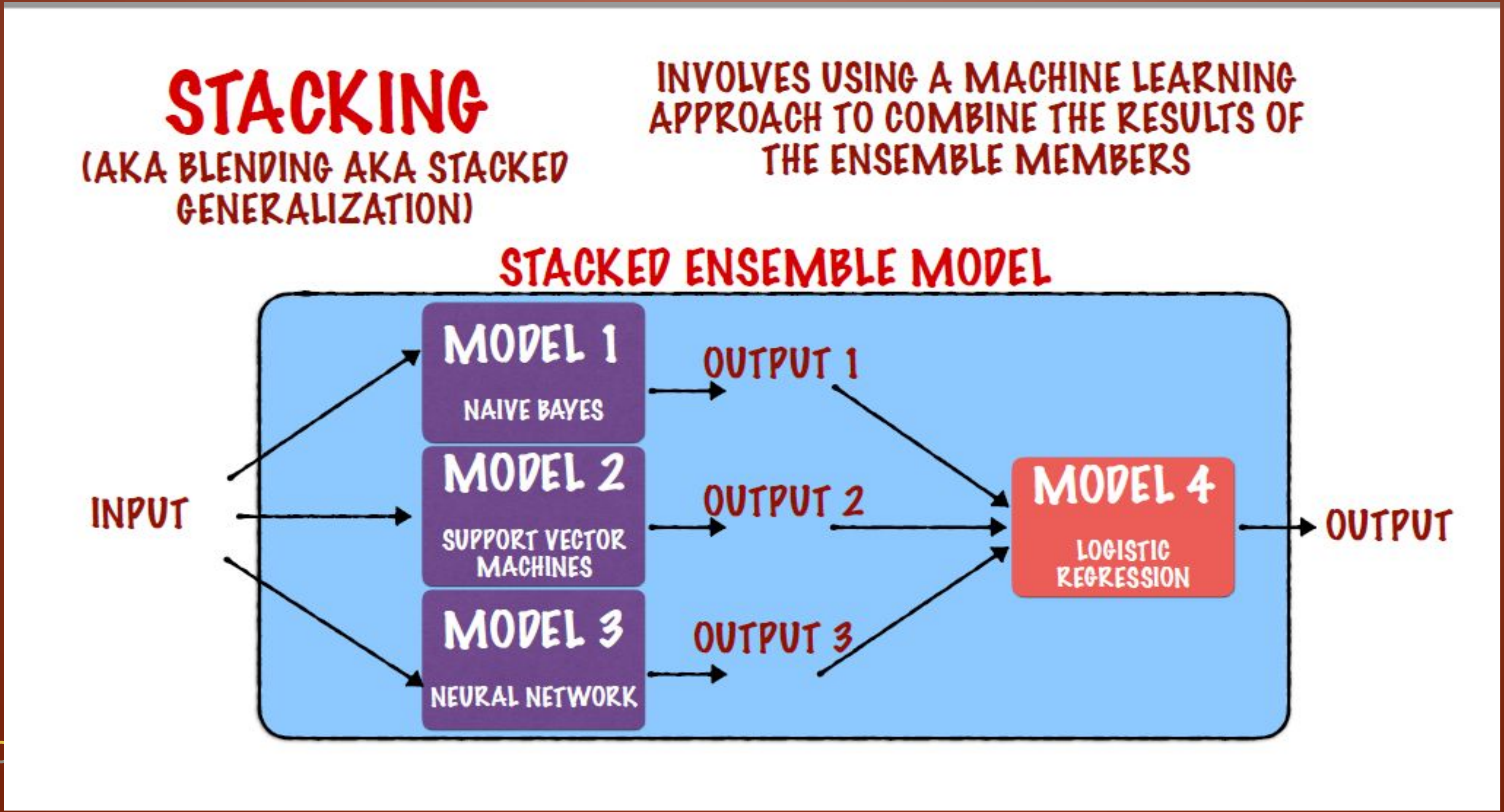
Intelligent Agent
Neuton

# RANDOM FOREST ALGORITHM

# **Bagging** (Ensemble Technique)

- Also known as Bootstrap Aggregation.
- Row Sampling with Replacement is performed (Bootstrap).
- Each Model trains and predicts an outcome and the majority is taken as final output (Aggregation).





# STACKING

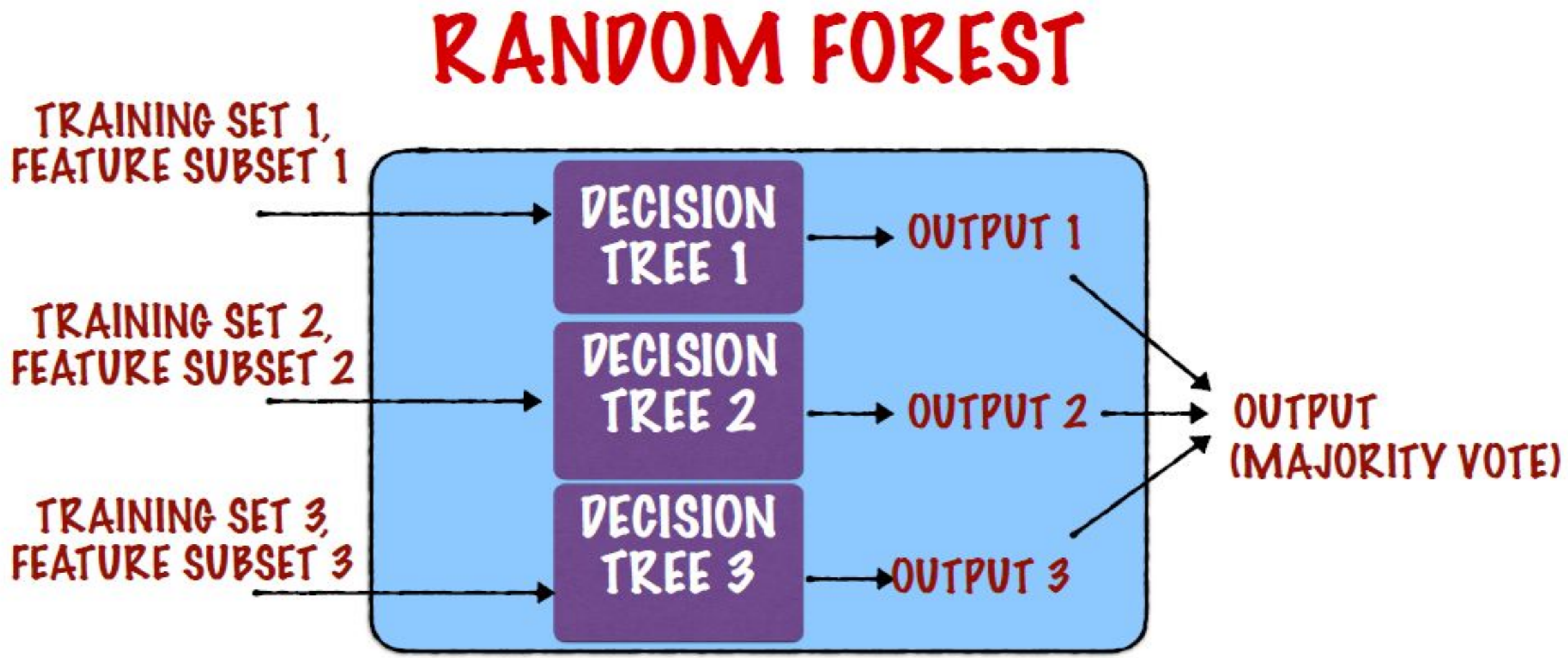
(AKA BLENDING AKA STACKED GENERALIZATION)

INVOLVES USING A MACHINE LEARNING APPROACH TO COMBINE THE RESULTS OF THE ENSEMBLE MEMBERS

## STACKED ENSEMBLE MODEL

# Random Forest

- Random forest, like it names implies consists of large number of individual decision trees that operate as an ensemble.
- Each individual tree has spits out a class prediction and class with most votes becomes our final model prediction.
- In Random Forest row sampling and feature sample is used which makes its a powerful algorithm.



RANDOM FOREST

TRAINING SET 1, FEATURE SUBSET 1

TRAINING SET 2, FEATURE SUBSET 2

TRAINING SET 3, FEATURE SUBSET 3

DECISION TREE 1

DECISION TREE 2

DECISION TREE 3

OUTPUT 1

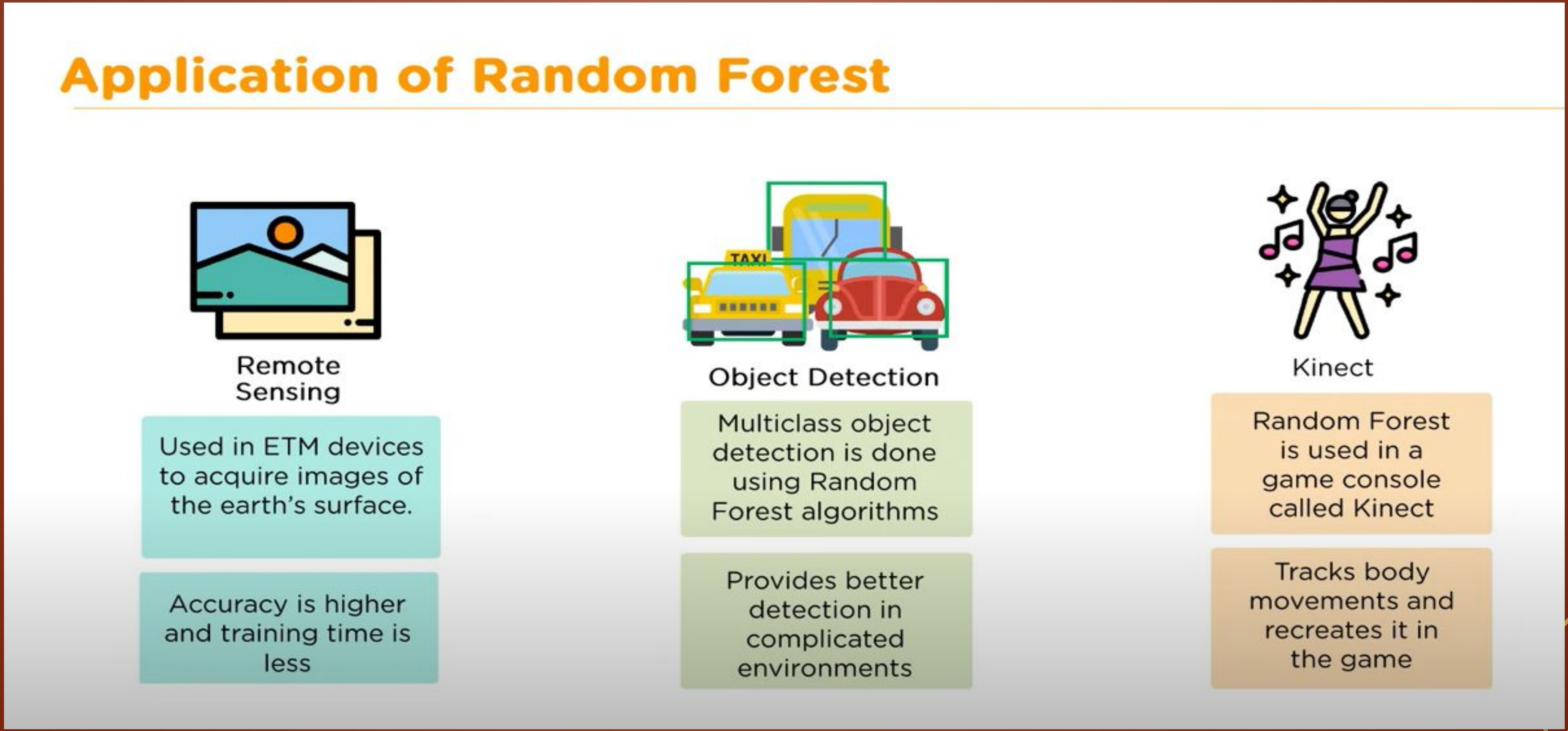
OUTPUT 2

OUTPUT 3

OUTPUT (MAJORITY VOTE)

Talent·Technology·Solutions Bellintegrator

Intelligent Agent Neuton



# Application of Random Forest

Remote Sensing

Used in ETM devices to acquire images of the earth's surface.

Accuracy is higher and training time is less

Object Detection

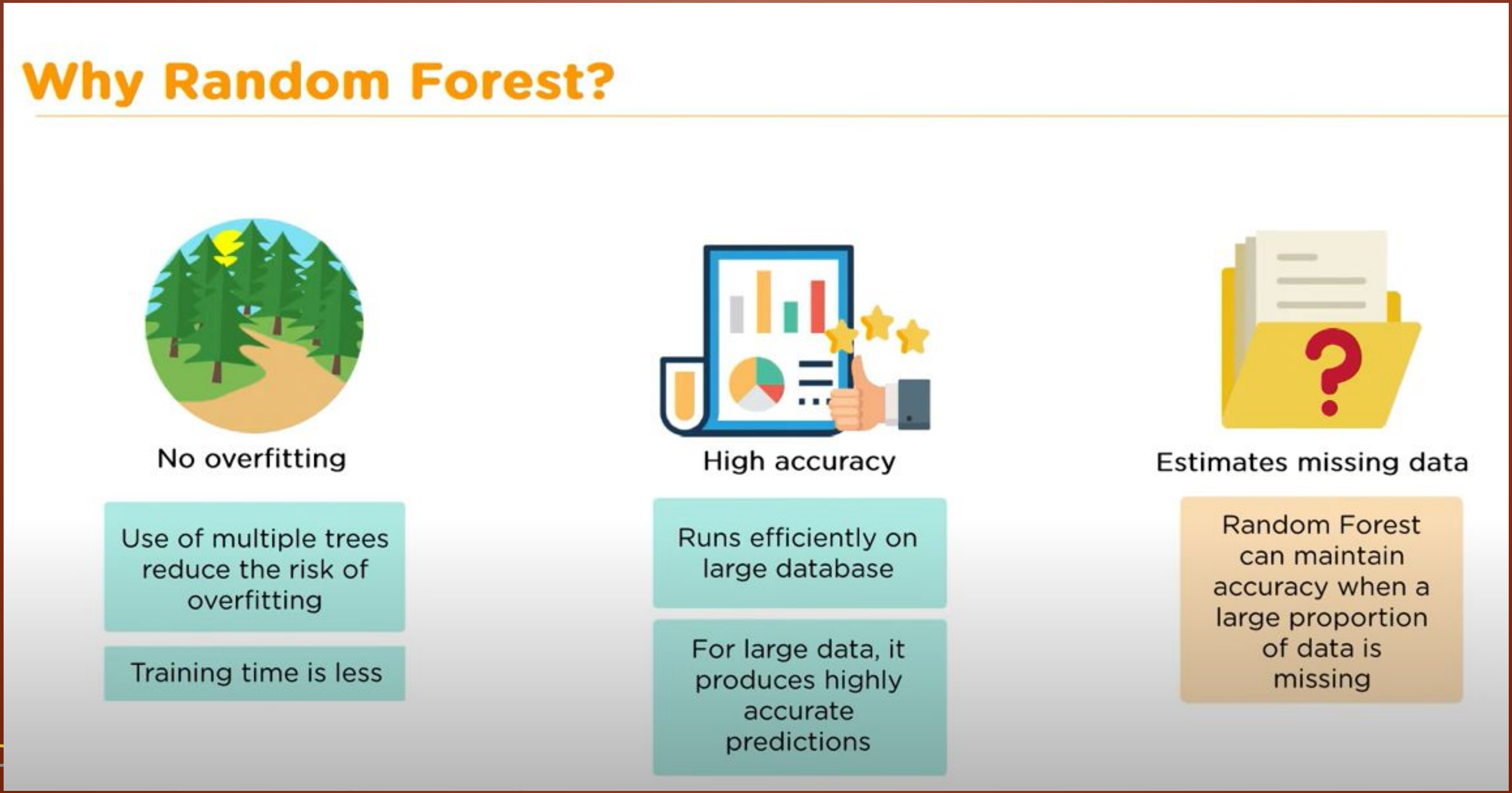
Multiclass object detection is done using Random Forest algorithms

Provides better detection in complicated environments

Kinect

Random Forest is used in a game console called Kinect

Tracks body movements and recreates it in the game



# Why Random Forest?

No overfitting

Use of multiple trees reduce the risk of overfitting

Training time is less

High accuracy

Runs efficiently on large database

For large data, it produces highly accurate predictions

Estimates missing data

Random Forest can maintain accuracy when a large proportion of data is missing

# Boosting (Ensemble technique)

- ADA Boost (Adaptative Boost)
- Gradient Boost
- XGB  Boost

# Performance Metrics

# ROC and AUC Curve

ROC : Receiver Operating Characteristic is mostly used to visualize binary classifier.

AUC : Area Under Curve is measure of ability of a classifier to distinguish and be the summary of ROC curve .

The ROC is plotted using the True Positive Rate and the False Positive Rate.

Talent · Technology · Solutions
Bellintegrator

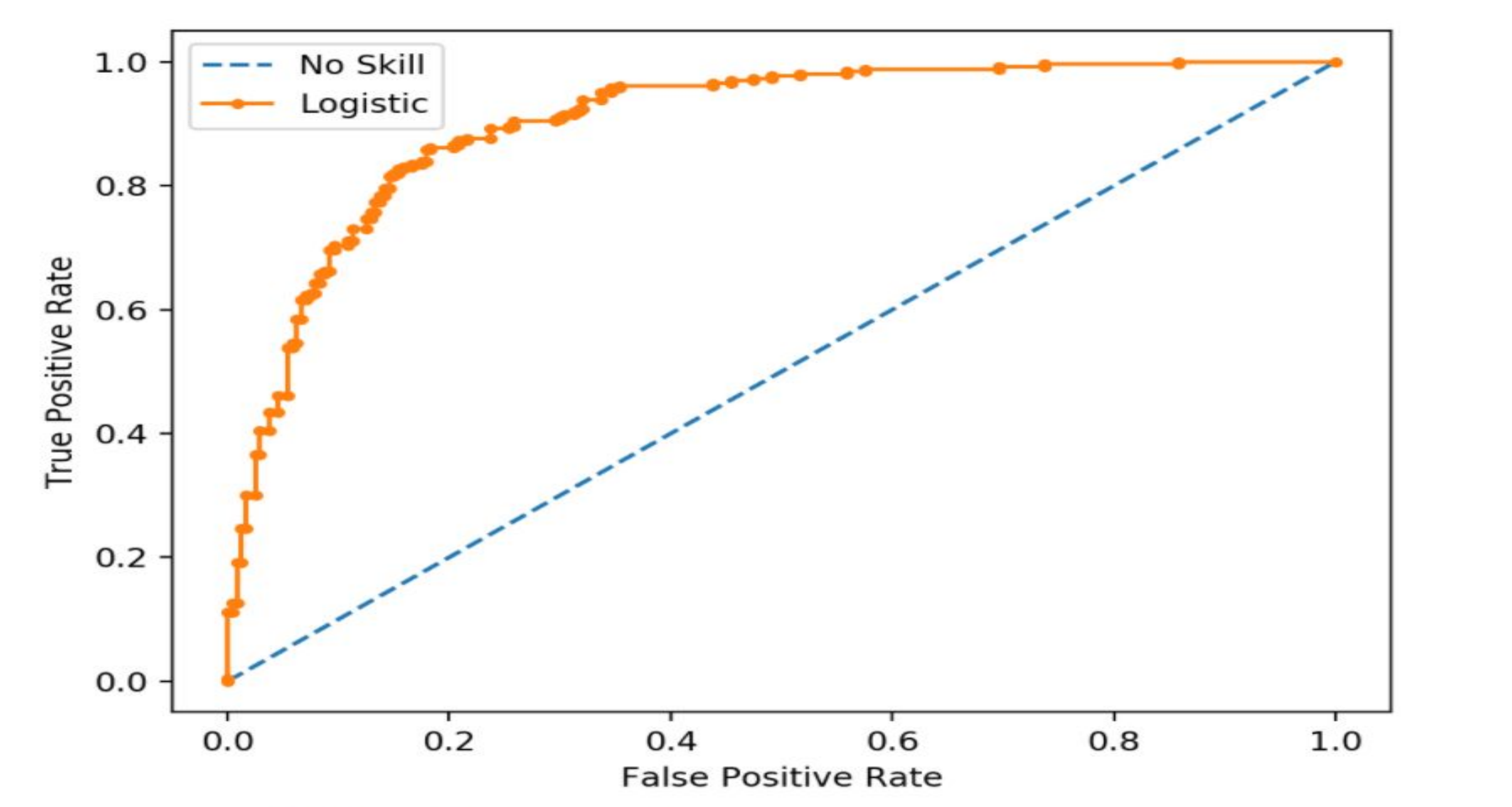
Intelligent Agent
Neuton

# ROC and AUC

- Its plotted is given as Sensitivity vs 1- Specificity.
- Sensitivity is True Positive Rate which is ratio of True positives and sum of True positives and False negatives.
- Specificity is False Positive Rate which is of False Positives and sum of True negatives and False positives.

Talent·Technology·Solutions
Bellintegrator

Intelligent Agent
Neuton



ROC Curve Plot for a No Skill Classifier and a Logistic Regression Model

# Industry Use cases of Decision Trees

- Commonly used in data mining .
- Loan Approval
- In Finance sector, forecasting future outcomes and assigning probabilities to those outcomes

# Industry Use cases of Random Forest

- Banking Sector : Banking sectors consists of most users. Used for Fraud Detection
- Medicines needs complex combination of chemicals. Random forest comnes handy in this scenario.
- Stock Market : Stock behaviour analysis can be done using Random Forest.

# Summary

- Decision Tree model is like a white box.
- Decision tree performs well in case of non linear data.
- Decision tree tends to overfit.
- Overfitting can be mitigated using various Ensemble techniques
- Random Forest is an ensemble of decision trees.
- Random forest widely popular because of its missing value handling and no overfitting.

Talent · Technology · Solutions
Bellintegrator

Intelligent Agent
Neuton

# THANK YOU

## Hope You Got To Learn Something!!!